

# Course 02402 Introduction to Statistics

## Lecture 7: Simulation-based statistics

DTU Compute  
Technical University of Denmark  
2800 Lyngby – Denmark

# Overview

- 1 Introduction to simulation - what is it really?
  - Example: Area of plates
- 2 Propagation of error
- 3 Parametric bootstrap
  - Introduction to bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval assuming any distributions
- 4 Non-parametric bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval

# Overview

- 1 Introduction to simulation - what is it really?
  - Example: Area of plates
- 2 Propagation of error
- 3 Parametric bootstrap
  - Introduction to bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval assuming any distributions
- 4 Non-parametric bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval

# Motivation

- Many (most?) relevant statistics (“computed features”) have complicated sampling distributions. One might want to do statistical inference for, e.g.:
  - The median
  - Quantiles in general, or perhaps  $IQR = Q_3 - Q_1$
  - The coefficient of variation
  - Any non-linear function of one or more input variables
  - (The standard deviation)
- The distribution of the data itself may be non-normal, complicating the statistical theory for even the simple mean.
- We may hope for the magic of the CLT (Central Limit Theorem).
- But: We never *really* know whether the CLT is good enough in a given situation - simulation can tell us!
- Requires: Use of a computer with software that can do simulations. R is a super tool for this!

# What is simulation really?

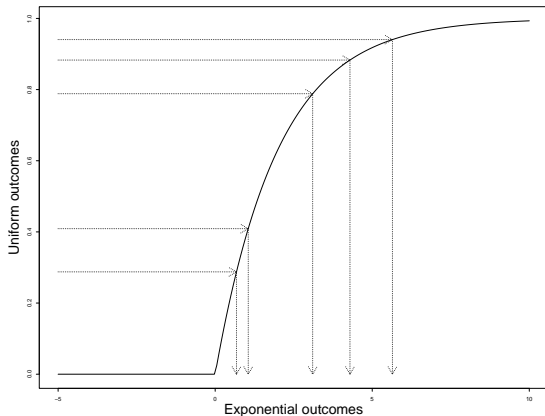
- (Pseudo) random numbers are generated using a computer.
- A random number generator is an algorithm that can generate  $x_{i+1}$  from  $x_i$ .
- The resulting sequence of numbers appears random.
- Requires a “starting point” called a *seed*.
- Basically, the uniform distribution is simulated in this manner, and then:

**Theorem 2.51:** All distributions can be extracted from the uniform

If  $U \sim \text{Uniform}(0, 1)$  and  $F$  is a distribution function for any probability distribution, then  $F^{-1}(U)$  follows the distribution given by  $F$ .

# Example: The exponential distribution with $\lambda = 0.5$ :

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0.5x}$$



## In practice, in R

Many distributions are ready for simulation, for instance:

---

<code>rbinom</code>	The binomial distribution
<code>rpois</code>	The Poisson distribution
<code>rhyper</code>	The hypergeometric distribution
<code>rnorm</code>	The normal distribution
<code>rlnorm</code>	The log-normal distributions
<code>rexp</code>	The exponential distribution
<code>runif</code>	The uniform distribution
<code>rt</code>	The t-distribution
<code>rchisq</code>	The $\chi^2$ -distribution
<code>rf</code>	The F-distribution

---

## Example: Area of plates

A company produces rectangular plates. The length of a plate (in meters),  $X$ , is assumed to follow a normal distribution  $N(2, 0.01^2)$ . The width of a plate (in meters),  $Y$ , is assumed to follow a normal distribution  $N(3, 0.02^2)$ . We are interested in the area of the plates, which is given by  $A = XY$ .

- What is the mean area?
- What is the standard deviation of the area?
- How often do such plates have an area that differs by more than  $0.1 \text{ m}^2$  from the targeted  $6 \text{ m}^2$ ?
- (The probability of other events?)
- Generally: What is the probability distribution of the random variable  $A$ ?



## Example: Area of plates, solution by simulation

```
k = 10000 # Number of simulations  
X = rnorm(k, 2, 0.01)  
Y = rnorm(k, 3, 0.02)  
A = X*Y
```

```
mean(A)
```

```
[1] 6
```

```
var(A)
```

```
[1] 0.002458
```

```
mean(abs(A - 6) > 0.1)
```

```
[1] 0.0439
```

# Overview

- 1 Introduction to simulation - what is it really?
  - Example: Area of plates
- 2 Propagation of error
- 3 Parametric bootstrap
  - Introduction to bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval assuming any distributions
- 4 Non-parametric bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval

# Propagation of error

Must be able to find:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$$

# Propagation of error

Must be able to find:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$$

We already know:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \sum_{i=1}^n a_i^2 \sigma_i^2 \quad \text{if} \quad f(X_1, \dots, X_n) = \sum_{i=1}^n a_i X_i \quad (\text{and independence})$$

# Propagation of error

Must be able to find:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$$

We already know:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \sum_{i=1}^n a_i^2 \sigma_i^2 \quad \text{if} \quad f(X_1, \dots, X_n) = \sum_{i=1}^n a_i X_i \quad (\text{and independence})$$

Method ??: For non-linear functions, if  $X_1, \dots, X_n$  are independent,

$$\sigma_{f(X_1, \dots, X_n)}^2 \approx \sum_{i=1}^n \left( \frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2$$

## Example: Area of plates (continued)

We used a simulation method in the first part of the example.

Now, given two specific measurements of  $X$  and  $Y$ ,  $x = 2.00$  m and  $y = 3.00$  m: What is the variance of  $A = XY$ , using the error propagation law?

## Example: Area of plates (continued)

The variances are:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ and } \sigma_2^2 = \text{Var}(Y) = 0.02^2$$

## Example: Area of plates (continued)

The variances are:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ and } \sigma_2^2 = \text{Var}(Y) = 0.02^2$$

The function and its derivatives are:

$$f(x, y) = xy, \quad \frac{\partial f}{\partial x} = y, \quad \frac{\partial f}{\partial y} = x$$



## Example: Area of plates (continued)

The variances are:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ and } \sigma_2^2 = \text{Var}(Y) = 0.02^2$$

The function and its derivatives are:

$$f(x, y) = xy, \quad \frac{\partial f}{\partial x} = y, \quad \frac{\partial f}{\partial y} = x$$

So the result becomes:

$$\begin{aligned} \text{Var}(A) &\approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_2^2 \\ &= y^2 \sigma_1^2 + x^2 \sigma_2^2 \\ &= 3.00^2 \cdot 0.01^2 + 2.00^2 \cdot 0.02^2 \\ &= 0.0025 \end{aligned}$$

# Propagation of error - by simulation

## Method ??: Error propagation by simulation

Assume that we have actual measurements  $x_1, \dots, x_n$  with known/assumed error variances  $\sigma_1^2, \dots, \sigma_n^2$ .

- 1 Simulate  $k$  outcomes of all  $n$  measurements from assumed error distributions, e.g.  $N(x_i, \sigma_i^2)$ :  $X_i^{(j)}$ ,  $j = 1 \dots, k$ .
- 2 Calculate the standard deviation directly as the observed standard deviation of the  $k$  simulated values of  $f$ :

$$s_{f(X_1, \dots, X_n)}^{\text{sim}} = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (f_j - \bar{f})^2}$$

where

$$f_j = f(X_1^{(j)}, \dots, X_n^{(j)})$$

## Example: Area of plates (continued)

Actually, in this example, one *could* deduce the variance of  $A$  theoretically:

$$\begin{aligned}
 \text{Var}(XY) &= E[(XY)^2] - [E(XY)]^2 \\
 &= E(X^2)E(Y^2) - E(X)^2E(Y)^2 \\
 &= [\text{Var}(X) + E(X)^2] [\text{Var}(Y) + E(Y)^2] - E(X)^2E(Y)^2 \\
 &= \text{Var}(X)\text{Var}(Y) + \text{Var}(X)E(Y)^2 + \text{Var}(Y)E(X)^2 \\
 &= 0.01^2 \times 0.02^2 + 0.01^2 \times 3^2 + 0.02^2 \times 2^2 \\
 &= 0.00000004 + 0.0009 + 0.0016 \\
 &= 0.00250004
 \end{aligned}$$

## Example: Area of plates (continued)

Three different approaches:

- 1 The simulation based approach.
- 2 A theoretical derivation.
- 3 The analytical, but approximate, error propagation method.

## Example: Area of plates (continued)

Three different approaches:

- 1 The simulation based approach.
- 2 A theoretical derivation.
- 3 The analytical, but approximate, error propagation method.

The simulation approach has a number of crucial advantages:

- 1 It offers a simple tool to compute many other quantities than just the standard deviation. (The theoretical derivations of these could be much more complicated than what was shown for the variance).
- 2 It offers a simple tool to use any other distributions than the normal, if we believe that they reflect reality better.
- 3 It does not rely on linear approximations of the true non-linear relations.

# Overview

- 1 Introduction to simulation - what is it really?
  - Example: Area of plates
- 2 Propagation of error
- 3 **Parametric bootstrap**
  - Introduction to bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval assuming any distributions
- 4 Non-parametric bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval

# Bootstrapping

Bootstrapping exists in two versions:

- 1 Parametric bootstrap: Simulate multiple samples from the assumed (and estimated) distribution.
- 2 Non-parametric bootstrap: Simulate multiple samples directly from the data.

## Example: Confidence interval for an exponential mean

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0



## Example: Confidence interval for an exponential mean

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data, we estimate

$$\hat{\mu} = \bar{x} = 26.08 \text{ and hence: } \hat{\lambda} = 1/26.08 = 0.03834356$$

## Example: Confidence interval for an exponential mean

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data, we estimate

$$\hat{\mu} = \bar{x} = 26.08 \text{ and hence: } \hat{\lambda} = 1/26.08 = 0.03834356$$

Our distributional assumption:

The waiting times come from an exponential distribution.

## Example: Confidence interval for an exponential mean

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data, we estimate

$$\hat{\mu} = \bar{x} = 26.08 \text{ and hence: } \hat{\lambda} = 1/26.08 = 0.03834356$$

Our distributional assumption:

The waiting times come from an exponential distribution.

What is the confidence interval for  $\mu$ ?

Based on previous knowledge in this course: We don't know!

## Example: Confidence interval for an exponential mean

```
# Number of simulations
k <- 100000

# Simulate 10 exponentials with the 'right' mean k times
sim_samples <- replicate(k, rexp(10, 1/26.08))

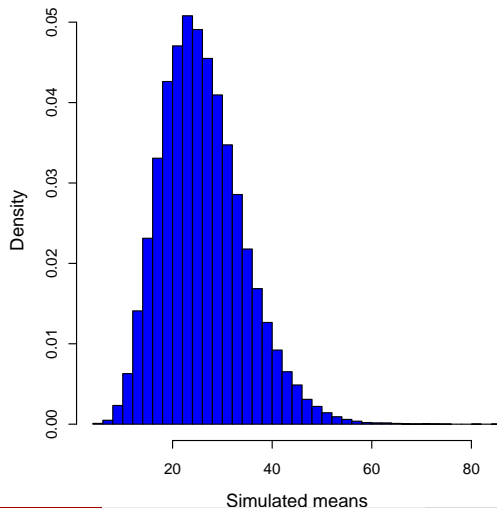
# Compute the mean of the 10 simulated observations k times
sim_means <- apply(sim_samples, 2, mean)

# Find relevant quantiles of the k simulated means
quantile(sim_means, c(0.025, 0.975))

## 2.5% 97.5%
## 12.59 44.63
```

# Example: Confidence interval for an exponential mean

```
# Make histogram of simulated means  
hist(sim_means, col = "blue", nclass = 30, main = "", prob = TRUE, xlab = "Simulated means")
```



## Example: Confidence interval for an exponential median

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

## Example: Confidence interval for an exponential median

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data we estimate

$$\text{Median} = 21.4 \text{ and } \hat{\mu} = \bar{x} = 26.08$$

## Example: Confidence interval for an exponential median

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data we estimate

$$\text{Median} = 21.4 \text{ and } \hat{\mu} = \bar{x} = 26.08$$

Our distributional assumption:

The waiting times come from an exponential distribution.



## Example: Confidence interval for an exponential median

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data we estimate

$$\text{Median} = 21.4 \text{ and } \hat{\mu} = \bar{x} = 26.08$$

Our distributional assumption:

The waiting times come from an exponential distribution.

What is the confidence interval for the median?

Based on previous knowledge in this course: We don't know!

## Example: Confidence interval for an exponential median

```
# Number of simulations
k <- 100000

# Simulate 10 exponentials with the 'right' mean k times
sim_samples <- replicate(k, rexp(10, 1/26.08))

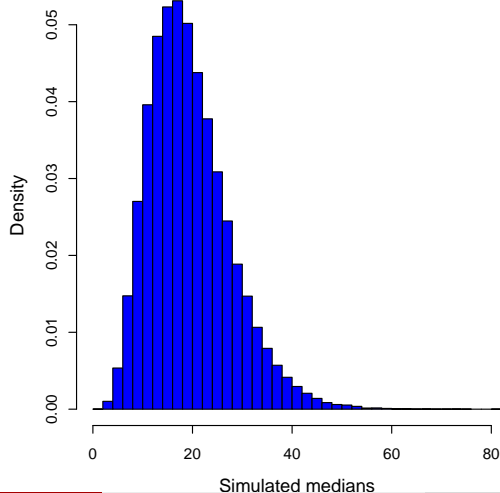
# Compute the median of the 10 simulated observations k times
sim_medians <- apply(sim_samples, 2, median)

# Find relevant quantiles of the k simulated medians
quantile(sim_medians, c(0.025, 0.975))

##    2.5%  97.5%
##  7.038 38.465
```

# Example: Confidence interval for an exponential median

```
# Make histogram of simulated medians  
hist(sim_medians, col = "blue", nclass = 30, main = "", prob = TRUE, xlab = "Simulated medians")
```



## Confidence interval for any feature (including $\mu$ )

### Method 4.7: Confidence interval for any feature $\theta$ by parametric bootstrap

Assume we have actual observations  $x_1, \dots, x_n$ , and that they come from some probability distribution with density  $f$ .

- 1 Simulate  $k$  samples of  $n$  observations from the assumed distribution  $f$  where the mean<sup>a</sup> is set to  $\bar{x}$ .
- 2 Calculate the statistic  $\hat{\theta}$  in each of the  $k$  samples to obtain  $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$ .
- 3 Find the  $100(\alpha/2)\%$  and  $100(1 - \alpha/2)\%$  quantiles of  $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$ ,  $q_{\alpha/2}^*$  and  $q_{1-\alpha/2}^*$ , to obtain the  $100(1 - \alpha)\%$  confidence interval:  
$$\left[ q_{\alpha/2}^*, q_{1-\alpha/2}^* \right]$$

---

<sup>a</sup>And otherwise chosen to match the data as well as possible: Some distributions have more than one mean related parameter, e.g. the normal or the log-normal. For these one should use a distribution with a variance that matches the sample variance of the data. Even more generally, the approach would be to match the chosen distribution to the data using the so-called *maximum likelihood* approach.

## Example: 99% CI for $Q_3$ assuming a normal distribution

```
# Heights data
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
n <- length(x)

# Define a Q3-function
Q3 <- function(x){ quantile(x, 0.75)}

# Set number of simulations
k <- 100000

# Simulate k samples of n = 10 normals with the 'right' mean and variance
sim_samples <- replicate(k, rnorm(n, mean(x), sd(x)))

# Compute the Q3 of the n = 10 simulated observations k times
simQ3s <- apply(sim_samples, 2, Q3)

# Find the two relevant quantiles of the k simulated Q3s
quantile(simQ3s, c(0.005, 0.995))

## 0.5% 99.5%
## 172.8 198.0
```

## Two-sample confidence interval for any feature comparison $\theta_1 - \theta_2$ (including $\mu_1 - \mu_2$ )

### Method 4.10: Two-sample confidence interval for any feature comparison $\theta_1 - \theta_2$ by parametric bootstrap

Assume we have actual observations  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$ , and that they stem from probability distributions with densities  $f_1$  and  $f_2$ .

- 1 Simulate  $k$  sets of 2 samples of  $n_1$  and  $n_2$  observations from the assumed distributions, setting the means<sup>a</sup> to  $\hat{\mu}_1 = \bar{x}$  and  $\hat{\mu}_2 = \bar{y}$ , respectively.
- 2 Calculate the difference between the features in each of the  $k$  samples:  $\hat{\theta}_{x1}^* - \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^* - \hat{\theta}_{yk}^*$ .
- 3 Find the  $100(\alpha/2)\%$  and  $100(1 - \alpha/2)\%$  quantiles for these,  $q_{\alpha/2}^*$  and  $q_{1-\alpha/2}^*$ , to obtain the  $100(1 - \alpha)\%$  confidence interval:

$$\left[ q_{\alpha/2}^*, q_{1-\alpha/2}^* \right]$$

## Example: Confidence interval for the difference of exponential means

```
# Day 1 data
x <- c(32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0)
n1 <- length(x)

# Day 2 data
y <- c(9.6, 22.2, 52.5, 12.6, 33.0, 15.2, 76.6, 36.3, 110.2,
       18.0, 62.4, 10.3)
n2 <- length(y)
```

## Example: Confidence interval for the difference of exponential means

```
# Set number of simulations:
k <- 100000

# Simulate k samples of each n1 = 10 and n2 = 12 exponentials
# with the 'right' means

simX_samples <- replicate(k, rexp(n1, 1/mean(x)))
simY_samples <- replicate(k, rexp(n2, 1/mean(y)))

# Compute the difference between the simulated means k times
sim_dif_means <- apply(simX_samples, 2, mean) -
  apply(simY_samples, 2, mean)

# Find the relevant quantiles of the k simulated differences of means:
quantile(sim_dif_means, c(0.025, 0.975))

## 2.5% 97.5%
## -40.74 14.12
```



# Parametric bootstrap - an overview

We assume *some* distribution!

Two confidence interval method boxes were given:

	One-sample	Two-sample
For any feature	Method 4.7	Method 4.10

# Overview

- 1 Introduction to simulation - what is it really?
  - Example: Area of plates
- 2 Propagation of error
- 3 Parametric bootstrap
  - Introduction to bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval assuming any distributions
- 4 **Non-parametric bootstrap**
  - **One-sample confidence interval for any feature**
  - **Two-sample confidence interval**

# Non-parametric bootstrap - an overview

We do *not* assume *any* distribution!

Two confidence interval method boxes will be given:

	One-sample	Two-sample
For any feature	Method 4.15	Method 4.17

## Example: Womens' cigarette consumption

In a study, womens' cigarette consumption before and after giving birth is explored. The following observations of the number of smoked cigarettes per day were obtained:

before	after	before	after
8	5	13	15
24	11	15	19
7	0	11	12
20	15	22	0
6	0	15	6
20	20		

Compare the before and after means! (Are they different?)

## Example: Womens' cigarette consumption

A paired  $t$ -test setting, *but* with clearly non-normal data!

```
# Data
x1 <- c(8, 24, 7, 20, 6, 20, 13, 15, 11, 22, 15)
x2 <- c(5, 11, 0, 15, 0, 20, 15, 19, 12, 0, 6)

# Compute differences
dif <- x1-x2
dif

## [1] 3 13 7 5 6 0 -2 -4 -1 22 9

# Compute average difference
mean(dif)

## [1] 5.273
```

# Example: Women's cigarette consumption - bootstrapping

```
t(replicate(5, sample(dif, replace = TRUE)))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## [1,]  -2   0   9  22   0  -1   0  -2   0   3   0
## [2,]  13   3  -2  -1  -2   7  13  -4  -2  -1   5
## [3,]   9  -4   5  -4   5   3  -4  13   3   0  22
## [4,]  -1  22  -2  -1  13   6  -4   0   0  -1  22
## [5,]   9  -2  13   6   9  22   0  -1   7   7  -1
```

## Example: Womens' cigarette consumption - the non-parametric results

Let us find the 95% confidence interval for the *mean* change in cigarette consumption.

```
k = 100000
sim_samples = replicate(k, sample(dif, replace = TRUE))
sim_means = apply(sim_samples, 2, mean)
quantile(sim_means, c(0.025,0.975))

## 2.5% 97.5%
## 1.364 9.818
```

# One-sample confidence interval for any feature $\theta$ (including $\mu$ )

## Method 4.15: Confidence interval for any feature $\theta$ by non-parametric bootstrap

Assume we have actual observations  $x_1, \dots, x_n$ .

- 1 Simulate  $k$  samples of size  $n$  by randomly sampling from the available data (with replacement).
- 2 Calculate the statistic  $\hat{\theta}$  for each of the  $k$  samples:  $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$ .
- 3 Find the  $100(\alpha/2)\%$  and  $100(1 - \alpha/2)\%$  quantiles for these,  $q_{\alpha/2}^*$  and  $q_{1-\alpha/2}^*$ , as the  $100(1 - \alpha)\%$  confidence interval:  $\left[ q_{\alpha/2}^*, q_{1-\alpha/2}^* \right]$



## Example: Womens' cigarette consumption

Let us find the 95% confidence interval for the *median* change in cigarette consumption in the example from above.

```
k = 100000
sim_samples = replicate(k, sample(dif, replace = TRUE))
sim_medians = apply(sim_samples, 2, median)
quantile(sim_medians, c(0.025,0.975))

## 2.5% 97.5%
## -1 9
```

## Example: Tooth health and infant bottle use

In a study, it was explored whether children who had received milk from a bottle had worse or better tooth health than those who had *not* received milk from a bottle. For 19 randomly selected children, it was recorded when they had had their first incident of caries:

bottle	age	bottle	age	bottle	age
no	9	no	10	yes	16
yes	14	no	8	yes	14
yes	15	no	6	yes	9
no	10	yes	12	no	12
no	12	yes	13	yes	12
no	6	no	20		
yes	19	yes	13		

## Example: Tooth health and infant bottle use - a 95% confidence interval for $\mu_1 - \mu_2$

```
# Reading in data
x <- c(9, 10, 12, 6, 10, 8, 6, 20, 12)
y <- c(14, 15, 19, 12, 13, 13, 16, 14, 9, 12)

# 95% CI for mean difference by non-parametric bootstrap
k <- 100000
simx_samples <- replicate(k, sample(x, replace = TRUE))
simy_samples <- replicate(k, sample(y, replace = TRUE))
sim_mean_difs <- apply(simx_samples, 2, mean) -
                  apply(simy_samples, 2, mean)
quantile(sim_mean_difs, c(0.025, 0.975))

##      2.5%    97.5%
## -6.2111 -0.1111
```

## Two-sample confidence interval for $\theta_1 - \theta_2$ (including $\mu_1 - \mu_2$ ) by non-parametric bootstrap

Method 4.17: Two-sample confidence interval for  $\theta_1 - \theta_2$  by non-parametric bootstrap

Assume we have actual observations  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$ .

- 1 Randomly draw  $k$  sets of 2 samples of  $n_1$  and  $n_2$  observations from the respective groups of data (with replacement).
- 2 Calculate the difference between the features in each of the  $k$  samples:  $\hat{\theta}_{x1}^* - \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^* - \hat{\theta}_{yk}^*$ .
- 3 Find the  $100(\alpha/2)\%$  and  $100(1 - \alpha/2)\%$  quantiles for these,  $q_{\alpha/2}^*$  and  $q_{1-\alpha/2}^*$ , to obtain the  $100(1 - \alpha)\%$  confidence interval:  
$$\left[ q_{\alpha/2}^*, q_{1-\alpha/2}^* \right]$$

## Example: Tooth health and infant bottle use - a 99% confidence interval for the difference of medians

```
k <- 100000
simx_samples <- replicate(k, sample(x, replace = TRUE))
simy_samples <- replicate(k, sample(y, replace = TRUE))
sim_median_difs <- apply(simx_samples, 2, median)-
                    apply(simy_samples, 2, median)
quantile(sim_median_difs, c(0.005,0.995))

## 0.5% 99.5%
## -8 0
```

# Bootstrapping - an overview

We were given 4 similar method boxes

- 1 With distribution assumptions or not (parametric or non-parametric).
- 2 For one- or two-sample analysis.

# Bootstrapping - an overview

We were given 4 similar method boxes

- 1 With distribution assumptions or not (parametric or non-parametric).
- 2 For one- or two-sample analysis.

Note:

*Means* also included in *other features*. Or: These methods may be used *not only* for means!

# Bootstrapping - an overview

We were given 4 similar method boxes

- 1 With distribution assumptions or not (parametric or non-parametric).
- 2 For one- or two-sample analysis.

Note:

*Means* also included in *other features*. Or: These methods may be used *not only* for means!

Hypothesis testing also possible

We can do hypothesis testing by looking at the confidence intervals!



# Overview

- 1 Introduction to simulation - what is it really?
  - Example: Area of plates
- 2 Propagation of error
- 3 Parametric bootstrap
  - Introduction to bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval assuming any distributions
- 4 Non-parametric bootstrap
  - One-sample confidence interval for any feature
  - Two-sample confidence interval